Une approche de fouille de textes pour l'identification automatique de relations spatiales

Sarah Zenasni, Eric Kergosien, Maguelonne Teisseire, Mathieu Roche

Une approche de fouille de textes pour l'identification automatique de relations spatiales

Sarah Zenasni[*,****] Eric Kergosein [**]
Maguelonne Teisseire [*,****] Mathieu Roche [*,****]

[*]UMR Tetis (Irstea, Cirad, AgroParisTech), France
prenom.nom@teledetection.fr
[**]GERIICO, Univ. Lille 3, France
prenom.nom@univ-lille3.fr
[****]LIRMM (CNRS, Université de Montpellier), France
prenom.nom@lirmm.fr

Résumé. La découverte de connaissances à partir de textes, en particulier l'identification d'informations spatiales, est une tâche difficile due à la complexité des textes écrits en langage naturel. Dans nos travaux, nous proposons une méthode combinant deux approches statistiques (analyse lexicale et contextuelle) et une approche de fouille de textes pour identifier les types de relations spatiales.

Contexte. Actuellement, les Systèmes d'Information Géographique (SIG) sont utilisés pour gérer les informations spatiales provenant de la cartographie ou de bases de données géographiques. Avec les informations aujourd'hui disponibles sur le Web toujours plus nombreuses (infobésité), de nouvelles sources de médias peuvent être exploitées pour extraire des informations géographiques (information spatiale, temporelle et thématique) à partir de données textuelles. Le travail présenté dans cet article se situe dans un tel contexte global, l'objectif étant de proposer une typologie de relations entre entités spatiales permettant d'identifier, finement et de manière automatique, des relations spatiales exprimées dans les textes. Globalement, les relations peuvent être identifiées par des calculs de similarité entre des contextes syntaxiques Grefenstette (1994), par prédiction à l'aide de réseaux bayésiens [Weissenbacher et Nazarenko (2007)], par des techniques de fouille de textes [Grčar et al. (2009)] ou encore par inférence de connaissances à l'aide d'algorithmes d'apprentissage [Giuliano et al. (2006)]. Ces méthodes sont efficaces, mais elles n'identifient pas toujours la sémantique de la relation. Nous proposons une méthode hybride, combinant deux approches statistiques (analyse lexicale et contextuelle) et une approche fouille de textes pour prédire les types de relations spatiales.

Approche hybride pour l'extraction de relations spatiales dans les textes. Dans un premier temps, nous avons adapté deux méthodes (String Matching [Maedche et Staab (2002)] et Lin [Lin (1998)]) afin de calculer la proximité lexicale entre les chaînes de caractères propres aux relations spatiales. Nous avons ensuite appliqué un algorithme de fouille de données (K plus proches voisins – KPPV [Bhatia (2010)]) pour prédire la classe des relations spatiales (région, direction, distance). Dans un second temps, nous proposons de considérer le contexte des relations (mots positionnés autour des relations) associé à l'algorithme KPPV pour prédire

les types de relations. Les mots du contexte sont pondérés sur la base de méthodes statistiques (fréquence, TF-IDF [Salton et Buckley (1988)], etc.). Enfin, pour tirer parties des deux types d'informations (lexicales et contextuelles), nous proposons une approche hybride qui est expérimentée sur des données réelles issues du challenge SemEval-2012 (corpus SPRL (Spatial Role Labeling) (Kordjamshidi et al., 2012)). Le corpus se compose de 1213 phrases annotées en anglais intégrant trois types de relations (région, direction, distance). Nous avons procédé à une série de tests dans lesquels nous avons fait varier les paramètres susceptibles d'influencer les résultats des mesures de performances (Exactitude, Précision, Rappel, F mesure), notamment le nombre de mots autour des relations dans l'analyse contextuelle et la valeur de K de l'algorithme KPPV. En terme de résultats, nous obtenons une Exactitude (Accuracy) de 0.84 selon une 3-validation croisée. Les résultats selon les meilleurs paramètres sont détaillés dans le tableau 1.

	Précision	Rappel	F-mesure
Région	0.87	0.94	0.90
Distance	0.75	0.42	0.54
Direction	0.82	0.79	0.80

TAB. 1 – *Résultats de combinaison*

Références

Bhatia, V. N. (2010). Survey of nearest neighbor techniques. *International Journal of Computer Science and Information Security 08*, 302–305.

Giuliano, C., A. Lavelli, et L. Romano (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of EACL*.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

Grčar, M., E. Klien, et B. Novak (2009). Using Term-Matching Algorithms for the Annotation of Geo-services. In *Knowl. Disc. Enhanced with Semantic and Social Information*, Chapter 8, pp. 127–143.

Kordjamshidi, P., S. Bethard, et M. Moens (2012). Semeval-2012 task 3: Spatial role labeling. In *Proceedings of SEM: Joint Conf. on Lexical and Computational Semantics*, pp. 365–373.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of ICML*, pp. 296–304.

Maedche, A. et S. Staab (2002). Measuring similarity between ontologies. In *Proceedings of EKAW*, pp. 251–263.

Salton, G. et C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage. 24*(5), 513–523.

Weissenbacher, D. et A. Nazarenko (2007). Identifier les pronoms anaphoriques et trouver leurs antécédents: l'intérêt de la classification bayésienne. In *Proceedings of TALN*, pp. 145–155.

Summary

Knowledge discovery from texts, particularly the identification of spatial information is a difficult task due to the complexity of the texts written in natural language. In our work, we propose a method combining two statistical approaches (lexical and contextual analysis) and a text mining approach to identify the types of spatial relationships.